NLP Pipeline

large-scale, parallel, multilingual and modularized

Multilingual text processing has never been so easy. Our Natural Language Processing pipeline is made of parallel, independent modules that make it possible to perform tasks like **language recognition**, **tokenization, morphological analysis**, **part-of-speech tagging**, **lemmatization** and **named entity recognition**.
In addition, our pipeline also features:

- **large scale**, making it possible to process millions of texts in seconds
- **parallelism** of independent modules, which can run dozens of tasks independently
- availability both as an **online service** and as an **offline software package**

Thanks to its integration with our companion products, WordAtlas®, Comprehendo® and Extraggo®, our pipeline also operates at a deeper level of analysis, making the following operations possible: **term**, **concept** and **entity extraction**, **domain labelling** and **word sense disambiguation** and **entity linking.**

LANGUAGES **SUPPORTED**

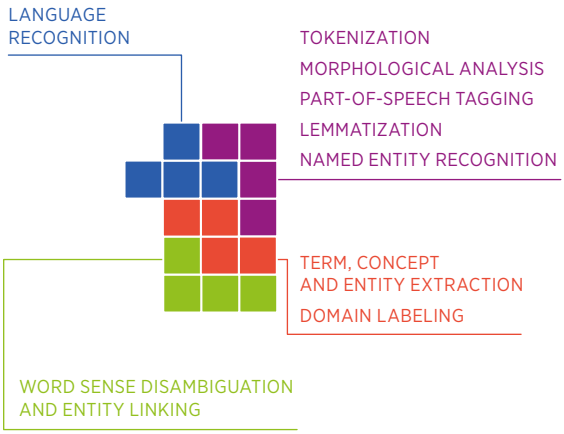
Our multilingual NLP Pipeline **fully supports 11 languages**:

AR Arabic	ZH Chinese	NL Dutch	EN English
FR French	DE German	IT Italian	PL Polish
PT Portuguese	ES Spanish	SV Swedish	

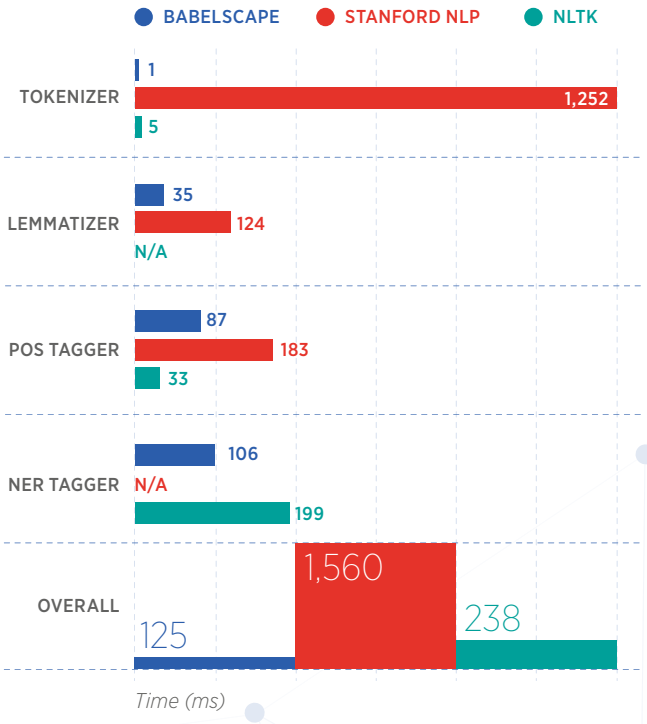
50 more languages are partially supported by a variable number of modules.

NLP PIPELINE **MODULES**

Babelscape's NLP pipeline includes modules which can be **accessed separately** and are integrated into the pipeline:



BABELSCAPE'S NLP PIPELINE **GOES FASTER!**



Babelscape
http://babelscape.com | Email: info@babelscape.com

How it works

The European Parliament is an important forum for political debates and decision-making.

Process

ENGLISH | 0.0260 sec | sentence 1

The	European	Parliament	is	an	important	forum	for	political	debates	and	decision-making	.
			be						*debate*			
DET	PROPN	PROPN	AUX	DET	ADJ	NOUN	ADP	ADJ	NOUN	CCONJ	NOUN	PUNCT
	ORG		third person singular present	singular		singular			plural		singular	

MODULES ALL OPERATE **SIMULTANEOUSLY**

LANGUAGE RECOGNITION

Babelscape's **language detector** identifies **60 languages**, including all European languages and most Asian languages.

ENGLISH

TOKENIZATION

Given a text, tokenization cuts it into minimal **meaningful semantic pieces** (tokens). The output tokens can then be input to subsequent modules for further processing.

The | European | Parliament | is | an | important | ...

PART-OF-SPEECH TAGGING

This module processes a text and **labels each word** with a **part-of-speech tag**.
In ambiguous cases - like when a word can be both a noun and a verb - it disambiguates the POS **relying on the context**.

The	European	Parliament	is	an	important	...
DET	PROPN	PROPN	AUX	DET	ADJ	

MORPHOLOGICAL ANALYSIS

The morpher performs full **morphological analysis** to provide detailed information about inflection, such as the tense of a verb or gender and number of a noun.

...	is	an	important	forum	...
	AUX	DET	ADJ	NOUN	
	third person singular present	singular		singular	

LEMMATIZATION

Lemmatization reduces the inflectional forms and sometimes derivationally related forms of a word to a **common base form**, i.e. its lemma.

... is an important forum for political debates ...
(to) be — debate

NAMED ENTITY RECOGNITION

NER seeks to locate and classify named entities mentioned in unstructured text into **pre-defined categories**, such as person names, organizations, locations, etc.

European Parliament
ORG

PER LOC

Babelscape

http://babelscape.com | Email: info@babelscape.com

MODULES THAT CAN BE **INTEGRATED** INTO THE PIPELINE

get the most out of your text processing by integrating our pipeline with other modules

WORD SENSE DISAMBIGUATION AND ENTITY LINKING

Identify the **proper meaning** of a dictionary word in a **given context** and link ambiguous words explicitly to **concepts and entities**.

Boost text analysis with **Comprehendo®**, the core engine of Babelscape's **multilingual text understanding**. This system works both with standard text and text snippets and brings two main advantages:

- It enables the **semantic aggregation** of similar information written using different words, either in the same language, or in **different languages**
- It enables **discrimination between meanings** of the same word

Thanks to its **semantic linkage** to **WordAtlas®**, text can be represented **independently of the source language** and **compared across languages**.

COMPREHENDO®

TERM, CONCEPT AND ENTITY EXTRACTION

Extract **key insights** from unstructured data independently of the languages used in the input. With **Extraggo®** - our state-of-the-art system for text processing - you can distill knowledge by:

- Extracting the **terms**, **key concepts** and **entities** involved in a text
- **Interconnecting** them in an intelligent way

DOMAIN LABELING

Get the core of any text and organize key concepts and domains with a seamless combination of **statistical and semantic techniques**. This module can:

- **Rank** - by importance - **terms**, **concepts** and **entities**
- Identify the **domains** of the text
- **Generalize** terms to **concepts and named entities**

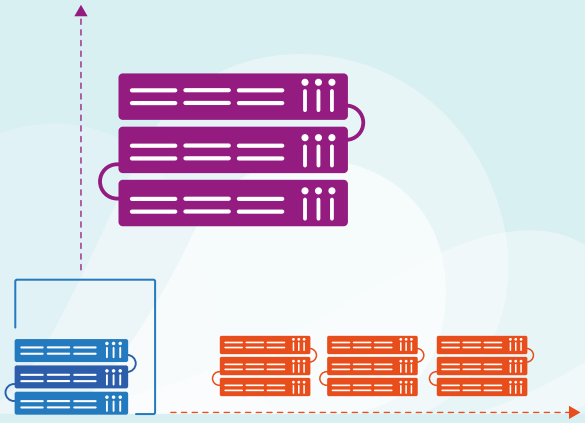
Extraggo® makes it possible to **distill knowledge from text written in multiple languages** and it lets you organize and manage knowledge easily.

EXTRAGGO®

Scalability

With our pipeline you can go from zero to infinity: its scalability will **embrace every need**.

We have developed our pipeline giving priority to the highest levels of **vertical and horizontal scalability**. Our pipeline is ready to be used on cloud infrastructures through the use of the most used orchestrators.



Pricing

Our pipeline has been conceived with a **modular pricing model**, so as to be a **flexible solution** to your specific business requirements.

You can choose the **number of languages**, go for **online** vs. **offline** usage and add **extra modules**. Every company is unique, and that's why we will create a **customized plan** for you. Contact us at info@babelscape.com.